5370_Reinforcelearning_assignment2

1. question desciption

> In a binomial model of a single stock with non-zero interest rate, assume that we can hedge any fraction of a stock, use policy gradient to train the optimal policy of hedging an ATM American put option with maturity T = 10. When do you early exercise the option? Is your solution same as what you obtain from delta hedging?

2. problem analyse

1. In the environment, we hold an American put option with strike price 100 of the single stock during the whole period.
2. The stock price is in a multi-step binomial model. The binomial model means the stock price can only move to two possible prices (Up or Down) with probability p(0.9 in this problem) and 1 − p respectively in each time step (day).
3. Each possible price filtration in each day is a single state.
4. Our target is to find an optimal policy to hedge the option, which means we expect to keep our wealth unchanged. In other words, our goal is to minimize （the hedging portfolio - payoff）^ 2
5. The stock pays no dividend; And there are no transaction costs and no taxes
6. For simplicity, there is no risk free rate and the discount rate (gamma) of reward is 1

3. project design

3.1 Environment

3.1.1 state:

The states are determined by time_t and price in time_t, in my experiment, I use number in binary system to indicate the path of price. For example, '101', means price down and up (the first 1 just show the price path is started).

3.1.2 reward:

Our target is to find an optimal policy to hedge the option, which means we expect to keep our wealth unchanged. Our goal is to minimize （the hedging portfolio - payoff）^ 2. In other words, to maximum -（the hedging portfolio - payoff）^ 2

3.2 Agent

3.2.1 action:

The agent have a continuous action space at each state (agent should choose his hedging strategy in each state)

For the policy gradient algorithm in this experiment, in order to fit in the problem with continuous action space, we assume that the mean and standard deviation for action taken in state s are respectively linear and linear-exponential in θ parameters

$$\mu(s,\theta) = \theta_\mu^T x_\mu(s)$$
$$\sigma(s,\theta) = exp(\theta_\sigma^T x_\sigma(s))$$

3.3 Algorithm

The agent applies policy gradient algorithm to improve its policy for hedging the put option. The method we used in this experiment is REINFORCE with Baseline.

REINFORCE with Baseline and Linear-Gaussian Policy: Monte-Carlo Policy-Gradient Control

Input: $\pi(a|s,\boldsymbol{\theta}) = \frac{1}{\sigma(s,\boldsymbol{\theta})\sqrt{2\pi}} \exp\left(-\frac{(a-\mu(s,\boldsymbol{\theta}))^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)$

where $\mu(s,\ \boldsymbol{\theta}) = \boldsymbol{\theta}_\mu^T \boldsymbol{x}_\mu(s)$ and $\sigma(s,\ \boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}_\sigma^T \boldsymbol{x}_\sigma(s)\right)$

Algorithm parameter: step size $\alpha > 0$, step size $\beta > 0$

Initialize policy parameter θ $(randomly)$ and rewards weights $w = 0$

Repeat forever:

Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(a|s,\boldsymbol{\theta})$

Loop for each step of the episode $t = 0, \ldots, T-1$:

$G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$

$w = w + \beta\gamma^t (G - w)$

$\boldsymbol{\theta}_\mu \leftarrow \boldsymbol{\theta}_\mu + \alpha\gamma^t (G-w) \nabla ln\pi\left(A_t|S_t,\boldsymbol{\theta}_\mu\right)$ where $\nabla ln\pi\left(A_t|S_t,\boldsymbol{\theta}_\mu\right) = \frac{1}{\sigma(s,\boldsymbol{\theta})^2}(a - \mu(s,\ \boldsymbol{\theta}))\boldsymbol{x}_\mu(s)$

$\boldsymbol{\theta}_\sigma \leftarrow \boldsymbol{\theta}_\sigma + \alpha\gamma^t (G-w) \nabla ln\pi(A_t|S_t,\boldsymbol{\theta}_\sigma)$ *where* $\nabla ln\pi(A_t|S_t,\boldsymbol{\theta}_\sigma) = \left(\frac{(a-\mu(s,\boldsymbol{\theta}))^2}{\sigma(s,\boldsymbol{\theta})^2} - 1\right)\boldsymbol{x}_\sigma(s)$

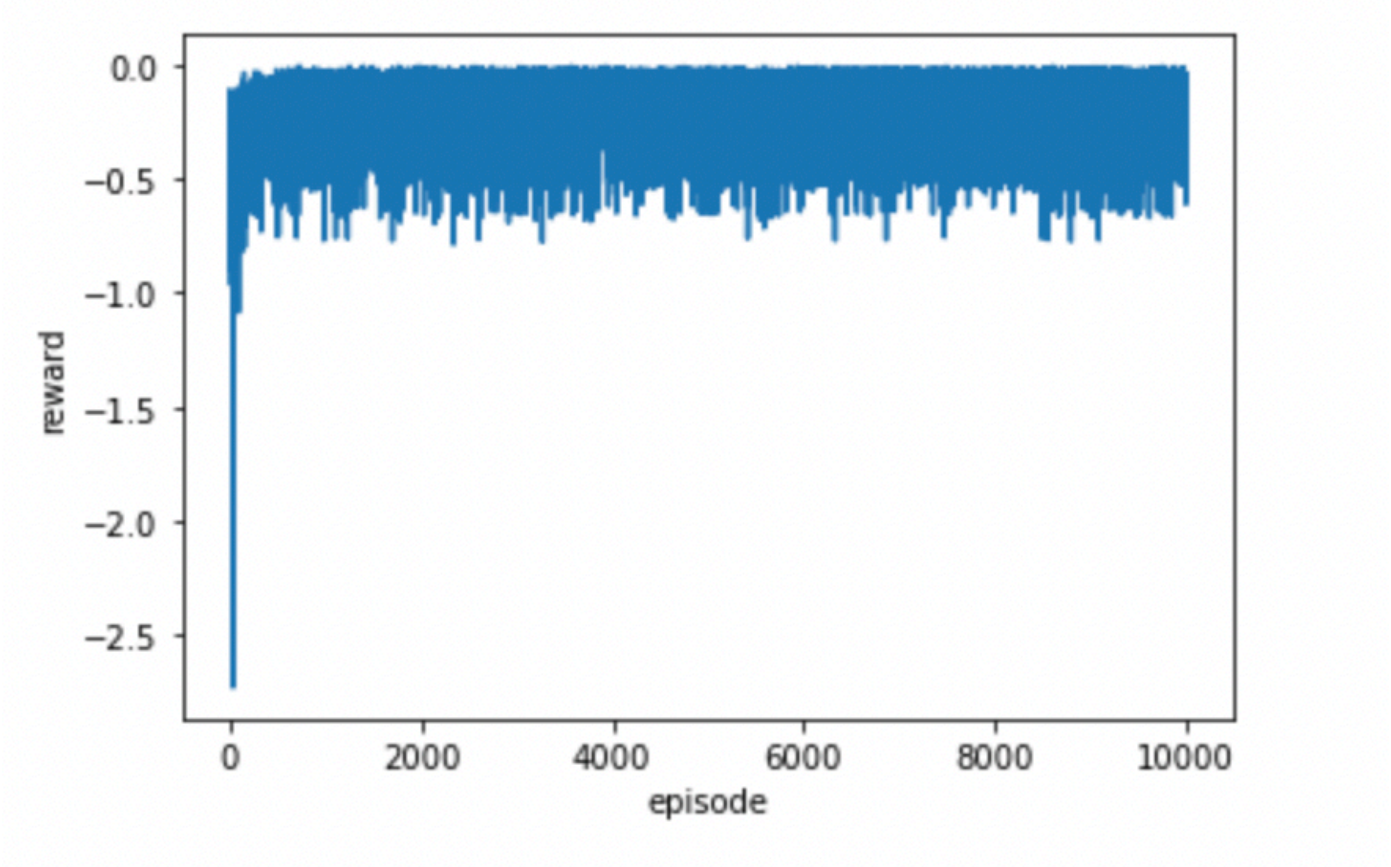
4. Delta Hedging policy

Delta hedging is to hedge option with Δ fraction of underlying stock. The option value in binomial model is:

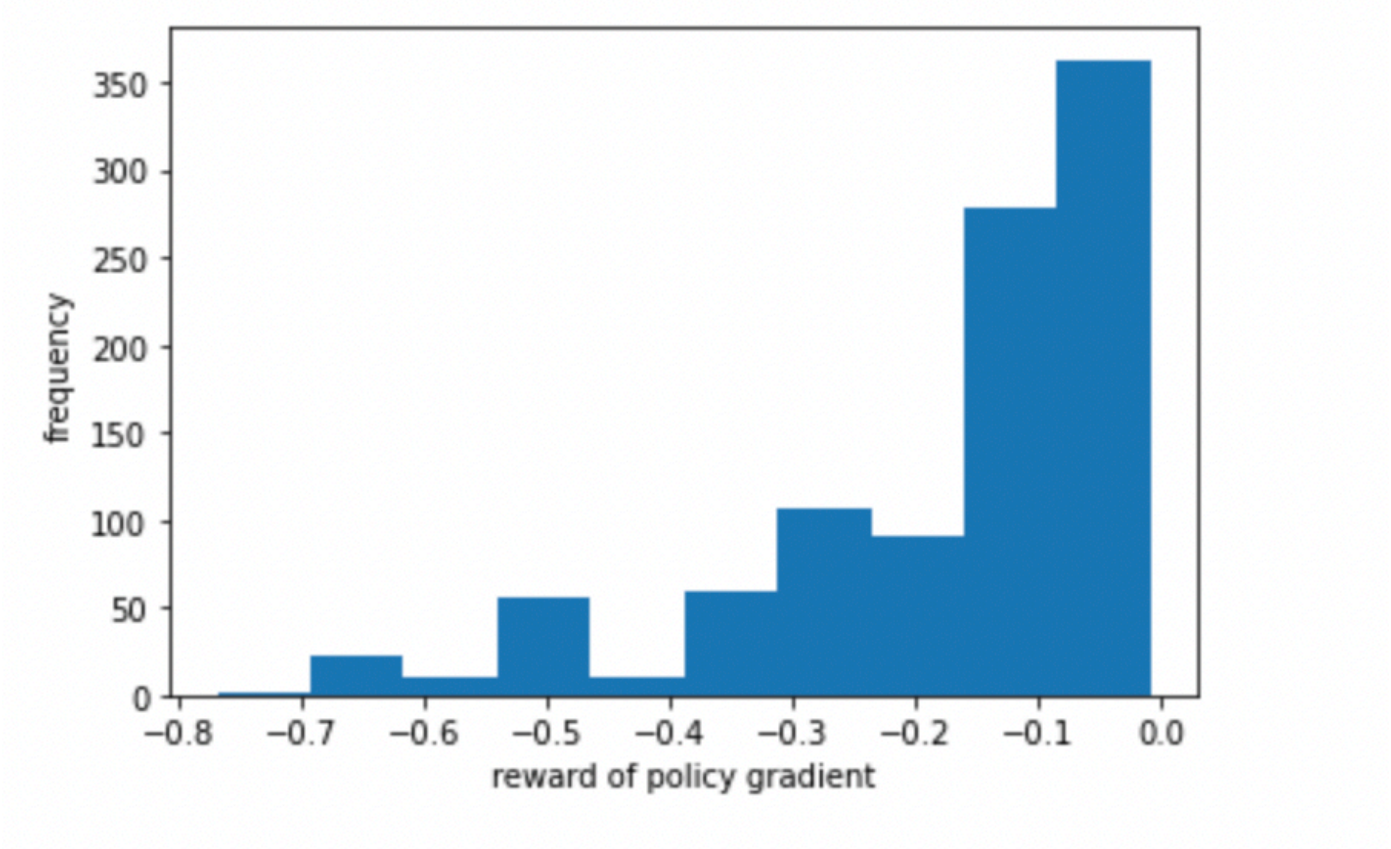
V = p * max(Su, K) + (1 − p) * max(Sd, K). And Δ = dV /dS

5.result analyse

The following figure shows the rewards for each episode. Obviously, while episodes increasing, the total reward of the agent becomes higher and asymptotically closer to its optimal value, zero. But since the environment is stochastic, the reward is unstable and fluctuates close to zero.

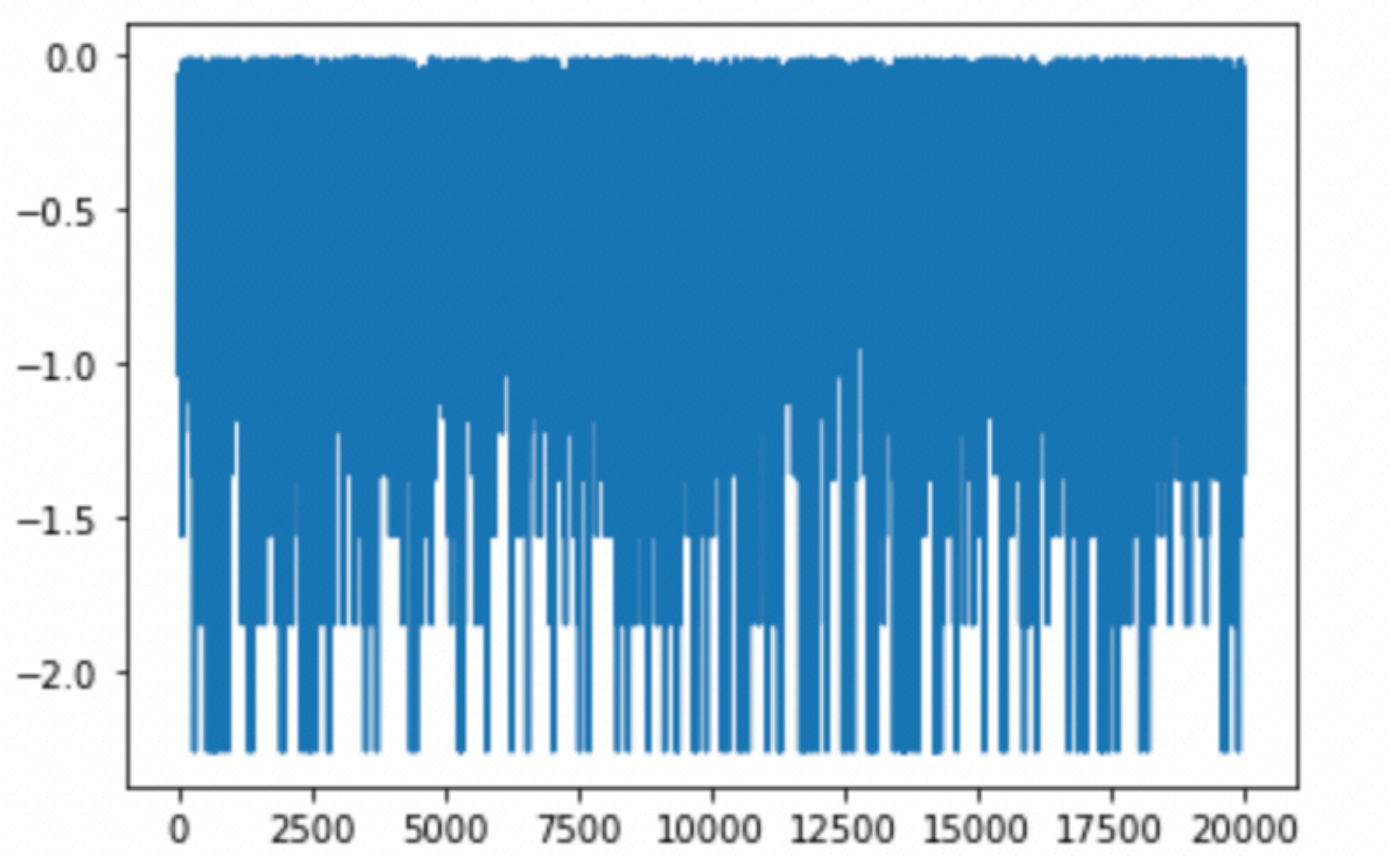


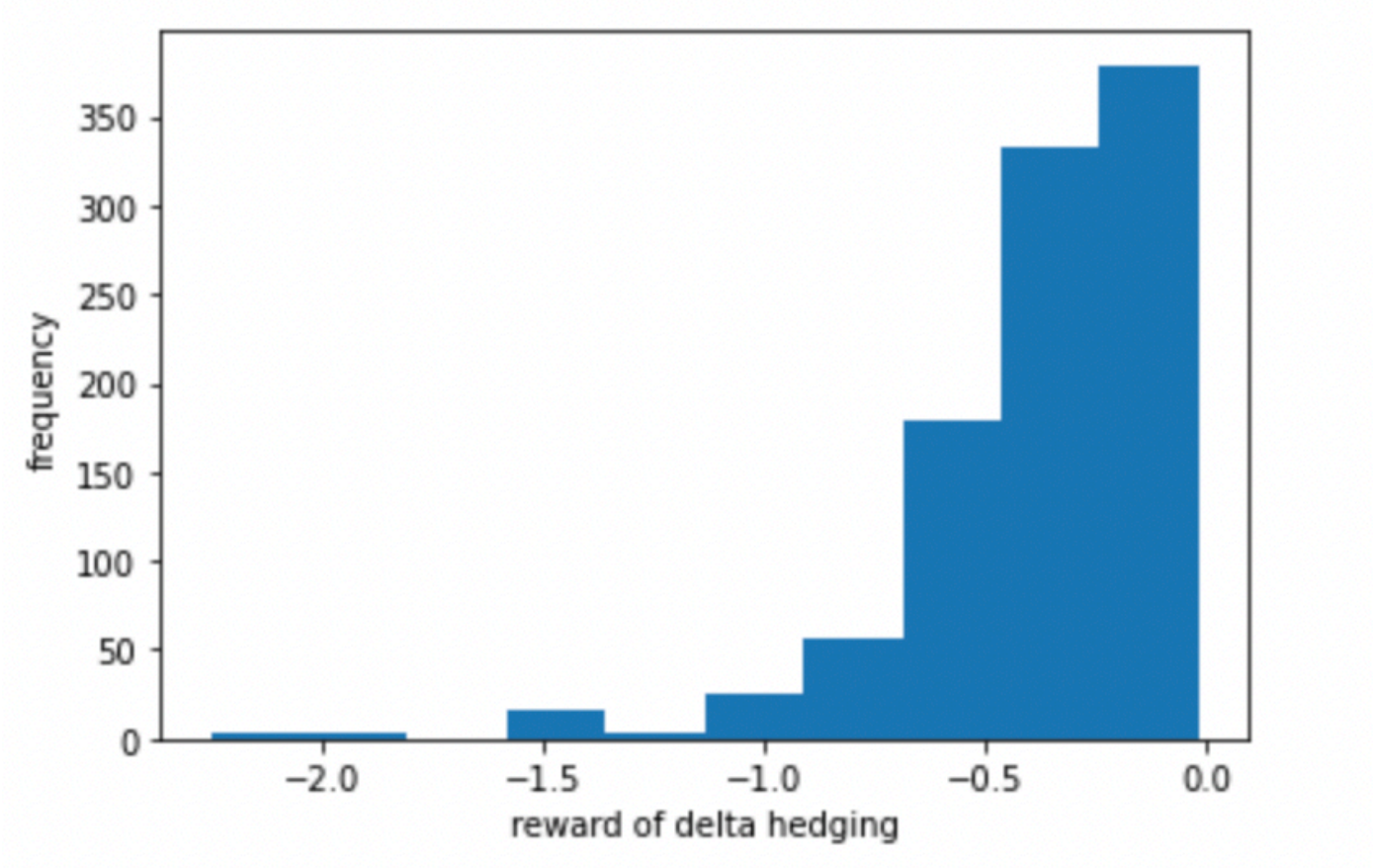
the following figure shows the reward distribution of policy gradient method.



5.2 compare to delta hedging results

The following figure shows the rewards for each episode and the reward distribution of delta hedging results .



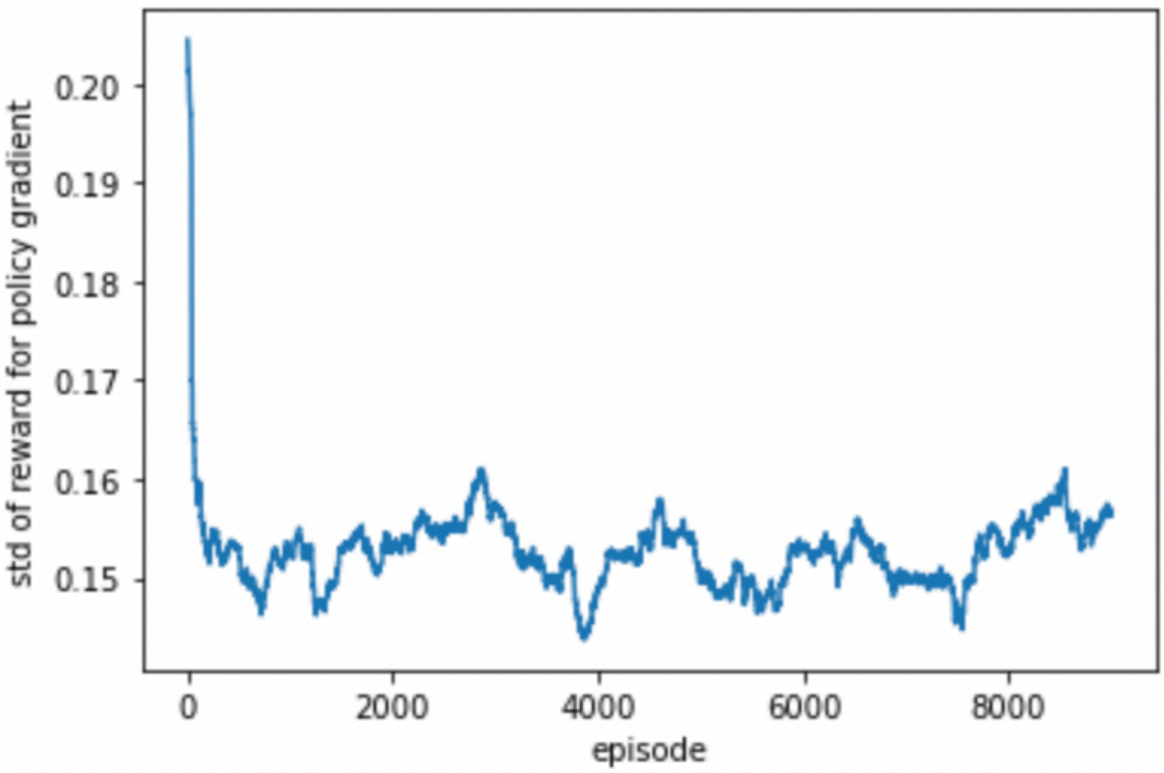


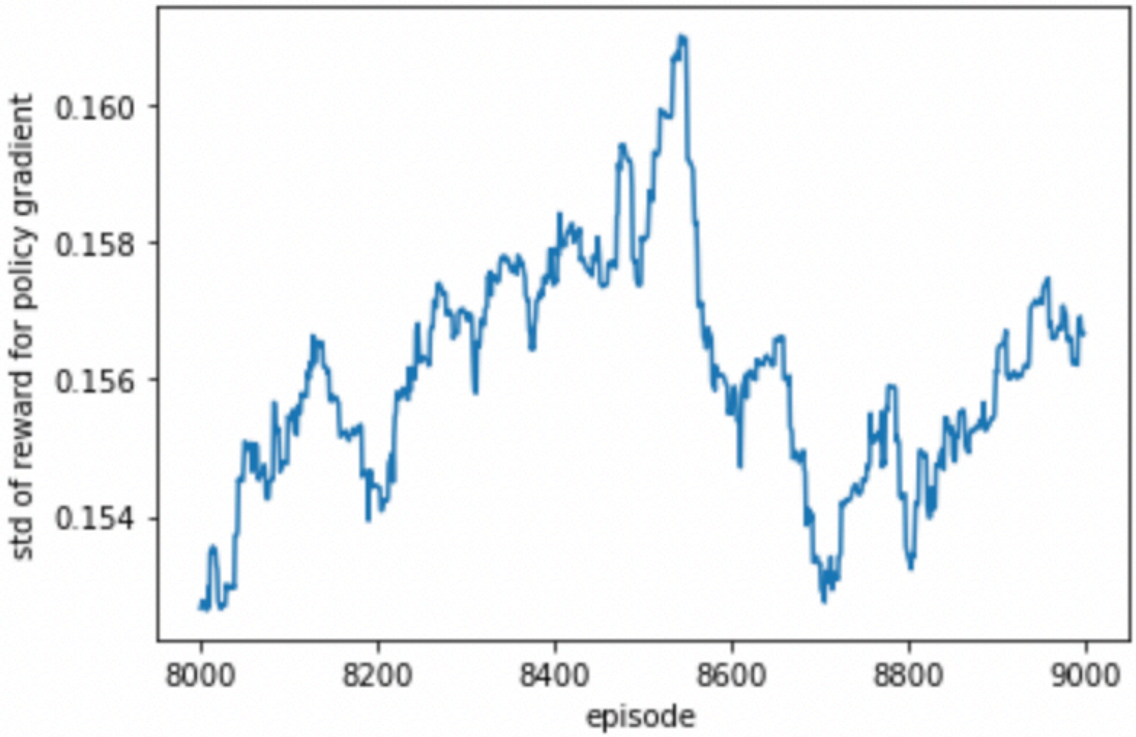
5.3 When do you early exercise the option

As a result of my experiment, we can find the reward function of each state. If we exercise the option will maximum the rewards, we should early exerciese the option.

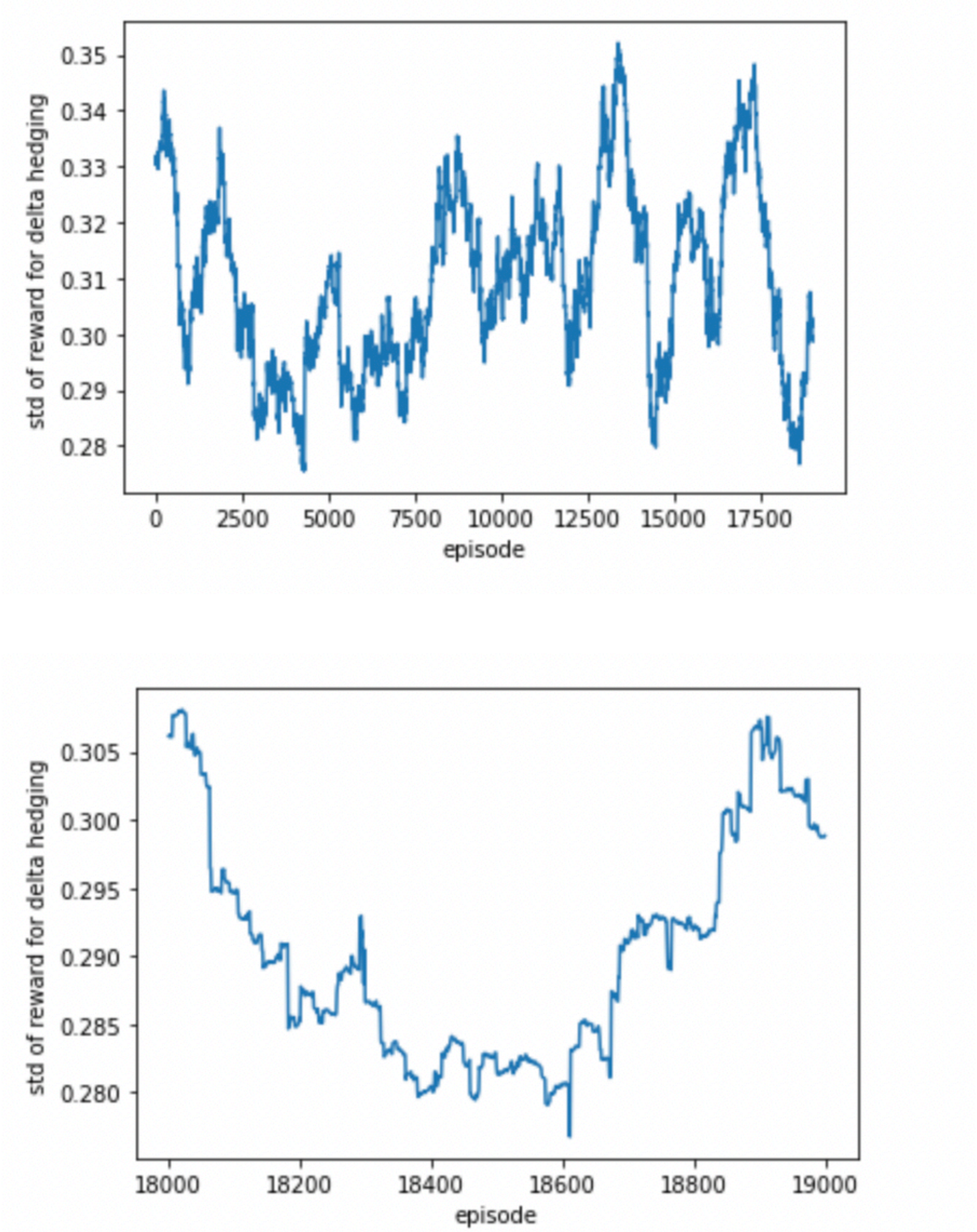
5.4 Standard Deviation of Reward

the following figures are about the standard deviation (Std) of rewards for different method.





std of rewards of policy gradient



std of rewards of delta hedging

We can see that Std of rewards for the policy gradient method will decrease and approach the Std of rewards for delta hedging method, while episode increasing.